DOCUMENT RESUME

ED 306 293 TM 013 163

AUTHOR Nitko, Anthony J.; Pettie, Allan
TITLE The Sixteen Quality Indicators: Standards for Evaluating Criterion-Referenced Tests.
PUB DATE Mar 89
NOTE 19p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, March 27-31, 1989). Text contains some small print.
PUB TYPE Speeches/Conference Papers (150) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Armed Forces; *Content Analysis; *Criterion Referenced Tests; Evaluation Methods; Formative Evaluation; Military Personnel; Quality Control; *Rating Scales; *Standards; Test Construction; *Test Reliability
IDENTIFIERS *Quality Indicators; *Sixteen Quality Indicators; Skill Qualification Test

ABSTRACT

The development, formative evaluation, and potential uses of the "Sixteen Quality Indicators" (16 QI) rating scale are described. The scale was developed as a systematic way to rate the quality of Skill Qualifications Tests (SQTs) in the United States Army. An SQT measures a soldier's knowledge of a military occupational specialty. It is a criterion-referenced test that samples the tasks in a specific specialty area. Several hundred SQTs are developed annually. The 16 QI is a list of critical criterion-referenced test characteristics. Scale drafts were reviewed by army job specialists and civilian testing experts to form a five-point scale. The 16 QI are grouped into characteristics of the total test, the task-measuring part of the test, and the item. The 16 QI rating scale has not yet been evaluated thoroughly, but would appear to have potential for monitoring SQT quality and diagnosing what needs to be done to improve the quality of a test. As an organized and systematic procedure, the 16 QI may be useful in other applications to evaluate criterion-referenced tests. Four tables present the elements of the 16 QI and the regulations and policies that support its use. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED306293

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Anthony J. Nitko

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

The Sixteen Quality Indicators:

Standards for Evaluating Criterion-Referenced Tests

by

Anthony J. Nitko
University of Pittsburgh

and

Allan Pettie
U.S. Army Training Support Center

A paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, California, March, 1989.

TM013168

BEST COPY AVAILABLE

Nitko
Pettie

The Sixteen Quality Indicators:

Standards for Evaluating Criterion-Referenced Tests

This paper describes the development, formative evaluation, and potential uses of the Sixteen Quality Indicators (16 QI) rating scale. The scale was developed as a systematic way to rate the quality of Skill Qualifications Tests (SQTs) in the U.S. Army. The concepts used in developing this rating scale may be useful to developing similar instruments for assessing the quality of criterion-referenced test development in other contexts.

Background

SQTs are one part of the U.S. Army's Individual Training Evaluation Program and its Enlisted Personnel Management System. An SQT measures a soldier's knowledge of a military occupational specialty (MOS). An MOS is a job classification (e.g., M48/M60 Armor Crewman). Soldiers must pass the SQT covering their MOS to maintain their certification. The domain of knowledge and abilities for an MOS is defined in detail by a Soldier's Manual which lists the tasks and performances (i.e., objectives) which comprise the MOS. An SQT is a criterion-referenced test that samples the tasks in a specific MOS domain. New forms of an SQT are developed for each MOS each year.

The Army Training Support Center (ATSC) provides guidance for the development of each SQT, but the actual development is the responsibility of one of the 21 proponent Army Training schools. Several hundred SQTs are developed each year and because of cost factors, the quality of only selected SQT is monitored. Guidance to the training schools' development staff is provided through test development regulations, specifically Regulation

351-2, Skill Qualification Test and Common Task Test Development Policy and Procedures.

In spite of the regulations, however, SQT quality is generally uneven. Implementation of SQT development principles varies widely from school-to-school, from MOS-to-MOS, and from one year's SQT to the next year's SQT. The regulations provide policy and guidance, but do not articulate a specific set of quality standards for systematically monitoring, in a relatively objective way, the quality of SQT scores. Without systematic monitoring it is difficult to (a) identify MOSs having better SQTs, (b) target special help to schools most in need of it, and (c) identify test development practices highly related to high quality SQTs.

One approach to this problem is to identify a small set of criterion-referenced test quality indicators and to organize these indicators into a standardized scale that can be used to systematically monitor the SQTs developed by each school. The set of indicators should meet psychometric validity and reliability criteria and be practical to use. Each quality indicator should be (a) related directly to the technical quality of the SQT scores which decision-makers use, (b) linked closely to existing policy, regulations, and accepted test development practices, and (c) of considerable diagnostic value for test developers who are charged with improving the tests.

Method of Developing the 16 QI

The authors have several years' experience in reviewing SQTs and working with SQT developers. Using this experience and suggestions for criterion-referenced test development in the psychometric literature, a list of critical criterion-referenced test characteristics were developed.

This list was refined by examining Regulation 351-2 and assuring that the quality indicators in the list were explicitly or implicitly implied by the official policy on SQT development. Scale drafts were circulated among ATSC staff members associated with SQT development and among civilian testing experts whom the Army had hired to review SQT quality in the recent past. The result was a list of 16 critical SQT characteristics which needed to be evaluated if the quality of an SQT was to be measured. Table 1 summarizes these characteristics.

The characteristics can be organized in several ways. One reviewer suggested organizing them by the categories: content adequacy, item-writing quality, and technical quality. This organization focuses on the nature of the expertise needed by a person to use the characteristics in evaluating an SQT. However, Table 1 shows the way chosen to organize them: quality of the total test scores, quality of the task (subtest) scores, and quality of the test items. This focuses evaluations of SQTs on the nature of the decisions which tend to be made from them. For example, a soldier must pass the total SQT with a minimum passing score of 60 on a standardized scale. Failing to pass places a soldier's MOS certification in jeopardy. Similarly, the regulations encourage individual soldier and group remediation of those who fail specific tasks' tests within an SQT. Task test scores are used for this purpose. Finally, since the subtest and total test scores are linked directly to the quality of the test items, it was deemed important to focus a good part of the evaluation on them.

The quality characteristics selected for the rating scale need to be justified not only on psychometric grounds but on policy grounds as well. A testing program is driven by the policy and decision context in which it will be used. Each of the characteristics selected for inclusion in Table 1 was

supported by some portion of the regulations pertaining to the SQT program. As an example, consider the second characteristic listed in Table 1, "decision consistency of the total score". The policy statements and regulations pay considerable attention to the minimum passing score and the use of SQT results to make pass-fail decisions. Table 2 illustrates how the regulations support the use of decision-consistency as one quality indicator of SQTs. Details of how each quality indicator is supported by Army policy and regulations are given elsewhere (Nitko, 1988).

Each quality indicator then needed to be oeprationalized before a scale could be formed. This required revieiwing the psychometric literature to identify recommended ways to measure or rate each quality characteristic. A number of indices for measuring decision-consistency , for example, have been presented in the literature (e.g., see Beck, 1984 and Subkoviak, 1984 for reviews). In this instance, Subkoviak's (1988) procedure for estimating Kappa coefficient, which uses coefficient alpha and a spcial table, was used because (a) coefficient alpha is a reasonably accurate indicator of an SQT's reliability, (b) this coefficient is calculated already by ATSC in connection with its item analysis report for each SQT, (c) the special tables provided by Subkoviak are relatively short and easy to use, and (d) only one administration of the test is needed. It should be noted that other investigators may have selected a different way to operationalize this quality indicator.

A third step was to translate the measure or index of a quality to a 5-point scale. This was needed in order to identify the quality levels of each indicator and to place each indicator on a similar quality scale. The quality scale, in turn, could communicate to test developers where each

SQT stood in relation to its quality rating on each indicator. Table 3 shows an example of this translation for the decision-consistency indicator. In this table, the "Excellent" or "4" category reflects Subkoviack's (1988) rule of thumb for judging the goodness of the Kappa coefficient. An alternate possibility for making this translation from measure to rating scale is to obtain distributions of the measure (e.g., Kappa coefficient) and use the quintiles of these distributions as break points for defining interval boundaries. This was not done for this version of the 16 QI.

Figure 1 shows the current version of the 16 QI rating scale. To the right of each verbal statement of the quality indicator is a horizontal bar marked in segments numbered 0 through 4. These numbered segments represent the quality ratings for that indicator. Below each bar are numbers which represent the interval boundaries of the quantified measure of that quality indicator. For example, for Quality Indicator 2, Decision-consistency of total score, the numbers below the bar represent values of Kappa coefficient. Thus a value of Kappa greater than or equal to .60 is given a rating of 4, .40 to .59 a 3, and so on.

The boundaries shown in this version of the 16 QI were set rationally using judgment and any guidance provided by Army SQT policy and suggestions from the psychometric literature. Both the index used for each quality indicator and the boundary for translating to quality ratings should be subject to further validation research.

Who Completes the 16 QI Rating Form

Although it is possible for one person to complete the 16 QI rating form, this is not necessary and may be undesirable. Different parts of the rating form require different kinds of competence to complete. Some parts

of the 16 QI are based on statistical analyses which already exist in or can be appended to the ATSC item analysis program (Indicators 2, 3, 6, 7, 8, and 14). The other quality indicators require reviewing and judging the quality of various aspects of an SQT. Subject-matter experts would be needed to judge the item-task congruence, whether items measure MOS-specific knowledge, and whether the keyed answer is correct. Testing specialists could judge the quality of the item-writing. Perhaps a team of persons could review several SQTs.

Possible Diagnostic Value of the 16 QI

One of the potential uses of the 16 QI is to point to specific ways in which an SQT could be improved. Since each quality indicator is operationally defined, a low rating implies that a specific test development action is needed to raise the rating. For example, to continue with Quality Indicator 2, decision-consistency, a low value of Kappa could be obtained because the test was too short (thus, lowering KR20) or because the minimum passing score needs to be adjusted. Table 4 lists each of the 16 QIs and gives suggestions as to how to raise a low rating on it.

Formative Evaluation and Current Status of the 16 QI

Because the 16 QI has not been evaluated thoroughly, it has no official status in the U.S. Army. It is currently undergoing formative evaluation so it may be improved. Empirical studies are under way to ascertain the extent to which the statistical indices for Indicators 2, 3, 6, 7, 8, and 14 are functioning to distinguish SQTs of various quality. Preliminary results indicate that the speededness index used for QI Number 3 is not distinguishing among different SQTs, even those which appear to be somewhat speeded. Also

the decision-consistency indices (Kappa coefficient) for task test scores (subtest scores) are quite low probably because many of the task tests are comprised of 4 to 7 items. Given that an SQT must cover 15 to 20 tasks, it may not be reasonable to insist that these subtests be made longer or, it may require that the Army not use these subtests to make individual training decisions at the task level. Also, Indicator 13, related to the distribution of answer patterns, seems not to distinguish SQTs. Apparently almost all current SQTs do not have a fixed or set pattern of correct answer choice positions. This raises the question of whether to keep 13 as a QI, even though it reflects the current regulations. If it were withdrawn from a quality monitoring instrument such as the 16 QI, violations of this rule might creep into the testing program (as it had in years past).

Some civilian testing specialists who are reviewing SQTs and who are using the 16 QI are uncomfortable judging Indicators 4 (item-task congruence) and Indicator 10 (whether items measure MOS specific knowledge), believing that a subject-matter expert should judge these qualities. Other civilian testing specialists seem not to mind doing this judging. A problem that arises here has to do with the nature of the SQT development effort. Subject-matter experts are usually noncommissioned officers who are assigned the job of writing and reviewing test items as a temporary assignment. They are not trained for the job and are often transferred after a short while. Thus, they frequently have no motivation to carefully review a test item to assure it exactly matches the task or that it cannot be answered by common sense, general knowledge, or other non-MOS specific means.

Another problem arose in connection with Quality Indicator 1, the extent to which an SQT represents the domain of tasks written in a Soldier's

Manual (SM). A SM covers all essential aspects of an MOS job. Previous Army regulations required that an SQT sample the entire domain implied by the SQT, preferably through stratified random sampling. Recently the regulation was changed so that SQT are to reflect only those tasks from the MOS which are considered necessary to make a soldier battle-ready. That is, each SQT is to be a purposive sample of tasks (perhaps all tasks) that will give it a "battle focus." Thus, the current QI on domain coverage is no longer valid.

Other studies which should be done before making the 16 QI operational include reliability and validity investigations. For example, several persons should independently rate the same SQTs using the 16 QI and the same data-base. The consistency among ratings should be studied. Further, several SQTs should be rated wholistically (perhaps by a team) and ranked according to perceived quality. Then, these same SQTs should be rated using the 16 QI. The two sets of ratings may be correlated to see if the 16 QT has some degree of predictive validity.

Summary

The 16 QI is a set of quality standards for systematically evaluating criterion-referenced tests developed in a decentralized testing program. The specific application discussed in this paper is the U.S. Army SQT testing program. The 16 QI has potential for monitoring sQT quality in this program. If specific SQTs consistently receive high ratings, this would indicate that the development process is probably working well. Consistently low ratings would indicate a breadwon in the developmental process and would signal the need to target technical assistance to specific SQT development units.

An important use of the 16 QI is in diagnosing what needs to be done

to improve the quality of a criterion-referenced test. Each of the 16 scales diagnoses a particular flaw in a test. Each flaw can be corrected by specific test development actions which will raise SQT quality. Table 4 described the actions a test developer should take to remediate a low rating on each quality indicator. Further, because the 16 QI is an organized and systematic rating procedure, one may easily monitor whether the remedial action has been taken and the impact it has had on test quality.

Although the 16 QI is presented in the context of the U.S. Army's SQT program, it has practical utility in other contexts. Many criterion-referenced programs are organized similarly to SQTs: domains are defined, domains are sampled, tests are designed to measure each sampled objective, and decisions about mastery are made for each objective and for the domain as a whole. With only slight modification, the 16 QI could be used to evaluate such criterion-referenced tests in other branches of the military, in occupational testing programs, and in public schools.

Finally, from a systems analysis perspective, the 16 QI could help identify criterion-referenced test development practices which consistently yield quality tests. Test quality may be measured by the 16 QI. An analysis of the test development process at a particular site can identify specific procedures which can be correlated with test quality indicators. Those procedures which consistently distinguish better tests from poor ones can be fostered at other test development sites.

References

Berk, R. A. (1984). Selecting the index of reliability. In R. A. Berk (Ed.). A guide to criterion-referenced test construction. Baltimore: Johns Hopkins University Press, pp. 231-266.

Brittain, C. V. (1987). Minimum passing score (MPS) on skill qualification tests (SQTs). (Memo dated 141402 July).

Nitko, A. J. (1988). The Sixteen Quality Indicators: A Rating Form for Evaluating Skill Qualification Tests. (Final Report). Fort Eustis, VA: Clay V. Brittain, U.S. Army Training Support Center (Contract DAAL03-86-D-001, Delivery Order 0534, Scientific Services Program).

Subkoviak, M. J. (1984). Estimating the reliability of mastery-nonmastery classifications. In R. A. Berk (Ed.), A guide to criterion-referenced test construction. Baltimore: Johns Hopkins University Press, pp. 267-291.

Subkoviak, M. J. (1988). Skill qualification test (SQT) and common task test (CTT) development policy and procedures. (RCS ATTG-17(R1). TRADOC Reg 351-2. Fort Monroe, VA: Headquarters, U.S. Army Training and Doctrine Command.

Table 1. Organization of the critical SQT characteristics which need to be assessed.

A. TOTAL TEST CHARACTERISTICS

1. SQT tasks as a representative sample of the SM domain
2. Decision-consistency of the total score
3. Sufficiency of testing time limits

B. TASK TEST CHARACTERISTICS

4. Congruence of items to task specifications
5. Inclusion of conditions of task performance on the test
6. Decision-consistency of task test scores
7. Length of task tests

C. ITEM CHARACTERISTICS

(a). Characteristics of items as functioning units

8. Easiness and difficulty of items
9. Performance-orientation of items
10. Items as measures of MOS-specific knowledge

(b). Characteristics of item stems

11. Freedom from flaws in phrasing the stem

(c). Characteristics of correct answers

12. Correctness of and freedom from ambiguity in the correct answer
13. Distribution of the correct answer position

(d). Characteristics of distractors

14. Plausibility of the distractors
15. Freedom from flaws in phrasing the distractors

(e). Other item characteristics

16. Freedom from other design flaws

Sixteen Quality Indicators for MOS Skill Qualification Tests

Evaluator _____ Date _____ SQT Test No _____

Quality Indicators — Ratings

1. Representativeness of SM domain

0	1	4
other	random	stratified

sampling plan used

2. Decision-consistency of total score

0	1	2	3	4
.00	.10	.20	.40	.60	1.00

Kappa coefficient

3. Sufficiency of testing time limits

0	1	2	3	4
	1.0	.9	.2	.1	0

Speededness Index

4. Task-item congruence

0	1	2	3	4
1.00	10	5	1	0

Percent items not matching tasks

5. Conditions of task performance

0	1	2	3	4
	4+	3	2	0

Number of tasks missing conditions

6. Decision-consistency of task test scores

0	1	2	3	4
.00	.10	.20	.40	.60	1.00

Average Kappa for task tests in SQT

7. Length of task tests

0	1	2	3	4
0.0	3.0	4.0	5.0	6.0+

Average number of items per task test

8. Easiness and difficulty of items

0	1	2	3	4
100	15	10	5	3	0

Percent of items that are too easy or too difficult

9. Performance-orientation of the test items

0	1	2	3	4
0	90	93	95	97	100

Percent of performance-oriented items

10. Items measuring MOS-specific knowledge

0	1	2	3	4
100	5	2	1	0

Percent of items not requiring MOS-specific knowledge

11. Phrasing the stems of items

0	1	2	3	4
100	15	10	5	3	0

Percent of items having flaws in the stems

12. Keyed answer correct and free from ambiguity

0	1	2	3	4
6+	5	3	1	0

Number of items miskeyed or have ambiguous answers

13. Distribution of correct answer positions

0	4
Discernable (set)	Not discernable (set)

Pattern of correct answers

14. Plausibility of distractors

0	1	2	3	4
100	15	10	5	3	0

Percent of items with fewer than 1% of lower group choosing a distractor

15. Phrasing the distractors of items

0	1	2	3	4
100	15	10	5	3	0

Percent of items with flaws in distractors

16. Other design characteristics of items which are not rated above

0	1	2	3	4
100	15	10	5	3	0

Percent of items having other design flaws

Summary of Quality Indicator Ratings

I. Total test score characteristics: Average of 1, 2, and 3 = _____

II. Task test score characteristics: Average of 4, 5, 6, and 7 = _____

III. Item characteristics: Average of 8, 9, 10, 11, 12, 13, 14, 15, and 16 = _____

IV. Overall SQT rating: Average of 1 through 16 = _____

Table 2. Examples of regulations and policy statements that support the need to use decision-consistency of the total score as a quality indicator for an SQT.

Statement/doctrine	Reference
a. SQT results indicate MOS proficiency for training and personnel management decisions	Reg. 351-2, Par 2-2b
b. SQTs are standardized so that decisions are consistent from one place and time to the next	ATSC, Bulletin 86-1, pg. 5
c. Minimum passing scores are to be set carefully and fairly	Brittain (1987)
d. Task test standards are set to maximize decision conssitency	Reg. 351-2, Par F-12g

Table 3. Example of the translation of a measure of a quality indicator to a quality rating. (In this case, translating the estimated Kappa coefficient for an SQT to a quality rating on a 5-point scale.)

Numerical value of Kappa for the SQT total test score	Rating Assigned	Possible interpretation
0.60 - 1.00	4	Excellent
0.40 - 0.59	3	Good
0.20 - 0.39	2	Mediocre
0.10 - 0.19	1	Poor
0.00 - 0.09	0	Very Poor

Table 4. What to do to raise a low rating on each area of the 16 QI Rating Form.

Quality Indicator	How to remediate a low rating
1. Representativesness of SM domain	1. Create and use a stratified random sampling plan for selecting tasks for the SQT
2. Decision-consistency of total score	2. (a) Increase the number of questions on the SQT (b) Adjust the MPS
3. Sufficiency of testing time limits	3. (a) Increase the SQT's time limits (b) Reduce the number of questions on the SQT (c) Make the SQT items less complicated
4. Task-item congruence	4. (a) Review each item carefully to be sure it matches the SM, TM, or FM task specifications (b) Use the murder board review process more effectively
5. Conditions of task performance	5. (a) Review and analyze more carefully the task descriptions found in the SM, TM, or FM (b) Create "situation" statements that capture the important task conditions
6. Decision-consistency of task test scores	6. (a) Increase the number of questions on these task tests with low decision-consistency coefficients (b) Eliminate from task test items that are too hard, too easy, or too complicated (c) Adjust the "go/no go" score
7. Length of task tests	7. Increase the average number of questions per task test
8. Easiness and difficulty of of items	8. (a) Rewrite difficult items to eliminate ambiguity, unnecessary complexity, and item-writing flaws (b) Replace "give-away", common sense, and copying items with performance-oriented items

Table 4 (continued)

9.	Performance-orientation of items	9.	(a) Be sure items require an actual performance of tasks where possible (b) Eliminate items asking for definitions of terms (c) Be sure items focus on who, what, where, when, how often, etc.
10.	Items measuring MOS specific knowledge	10.	(a) Eliminate items testing general knowledge, common sense, copy skills, simple reading skills (b) Write items that only those who can perform well on an MOS can answer correctly (c) Increase the ratio of "key" performances tested relative to the "essential" performances tested
11.	Phrasing the stems	11.	(a) Use standard testing and measurement guidelines and checklists to review and revise the item stems (b) Be sure the item stem is focused on a single performance and asks a direct question
12.	Keyed answer correct	12.	(a) Check the answer key before submitting to ATSC (b) Make more effective use of the murder board reviewers by asking them to actually take the SQT without seeing the answer key (c) Use the ATSC Expanded Item Analysis Report to identify items exhibiting ambiguous answers, then revise these items before using them again
13.	Distribution of correct answer positions	13.	(a) Review the SQT answer key to be sure there is no set pattern of keyed answers (b) When writing each item, put the response choices in a logical order
14.	Plausibility of distractors	14.	(a) Use the ATSC expanded Analysis Report to identify items exhibiting this flaw before using the item again (b) Eliminate non-functioning distractors (c) Replace nonfunctioning distractors with distractors based on

Table 4 (continued)

		errors or misconceptions of who are known to be among the poorest performers of that MOS (d) Administer stems without distractors to MOS holders: Use their responses as a basis for writing distractors
15.	Phrasing the distractors of items	15. Use standard testing and measurement sources and checklists to review each distractor set and correct the flaws identified
16.	Other design characteristics of items	16. (a) Follow the suggestions found in Regulation 351-2 for writing items and using pictorial material (b) Ask the murder board to review the items in light of the item-writing suggestions found in Reg 351-2